# Explore Dataset Structure

## Contents

# Structure of dataset

The PDS dataset consists of several data files. These contain data items defined by variables. Variables are taken from the eCRFs from Enroll-HD, REGISTRY, and Ad Hoc data. Specific data items have been transformed or obscured due to de-identification reasons.

*Studies within the combined PDS dataset*

All individuals in the PDS6 are Enroll-HD participants. However, PDS6 contains data gathered not only from Enroll-HD, but also integrates data from Registry, as well as Ad Hoc data. Study specific protocols and annotated eCRFs are housed under the *General Documents* section.

*Table 1: Data sources within Enroll-HD periodic dataset releases.*

| Study Name | Acronym | Chronological order |
|---|---|---|
| *Enroll-HD* | ENR | Enroll-HD is the most recent study a participant will have enrolled in. Participation in this study is mandatory for inclusion in the PDS |
| *REGISTRY V3* | R3 | Participation in Registry v3 is optional. Participation in this study precedes Enroll-HD |
| *REGISTRY V2* | R2 | Participation in Registry v2 is optional. Participation in this study precedes Enroll-HD, and Registry 3 (if available) |
| *Ad Hoc* | RET | Ad Hoc data are optional. These data are drawn from a variety of different sources, principally comprise UHDRS data, and are typically gathered prior to a participant's enrollment into Enroll-HD. Data are not collected under a specific study protocol. |

### *Data files within the combined PDS dataset*

Enroll-HD PDS releases are comprised of 11 data files, each of which fall into one of three categories:

**Participant-based**: *profile*, *pharmacotx*, *nonpharmacotx*, *nutsuppl, comorbid*

These contain general study-independent information about the participant. This information is applicable to all studies.

**Study-based:** *participation*, *events*

These contain study specific information about a participant within a study. Note the data file *events* is a special data file for Enroll-HD which contains all the reportable events of a participant.

**Visit-based:** *enroll*, *registry*, *adhoc, assessment*

These contain all visit-dependent information for the study, combined into one data file.

Each PDS data file is described in the table below.

*Table 2: Enroll-HD periodic dataset data file descriptions.*

| Data file | Type | Studies | Description |
|---|---|---|---|
| *profile* | participant | ENR, R3, R2, RET | General participant related data, some of which may be annually updated. This includes: Demographics, HDCC (HD clinical characteristics), CAG, Mortality |
| *pharmacotx* | participant | ENR, R3, R2, RET | Information about pharmacological therapies |
| *nutsuppl* | participant | ENR, R3, R2, RET | Information about nutritional supplements |
| *nonpharmacotx* | participant | ENR, R3, R2, RET | Information about nonpharmacological therapies |
| *comorbid* | participant | ENR, R3, R2, RET | Information about comorbid conditions and surgeries |
| *participation* | study | ENR, R3, R2, RET | Provides study specific information about study participation. Includes participant identifier, participant status, study start day, study end day (in event of participant withdrawal or death). Note: If a participant is enrolled in several studies, one line per study is provided |
| *event* | study | ENR | Enroll-HD study reportable event information |
| *assessment* | visit | ENR, R3, R2, RET | Visit-specific information about which assessments were performed at each visit (per study) |
| *enroll* | visit | ENR | Data from the Enroll-HD study |
| *registry* | visit | R3, R2 | Data from both the REGISTRY3 and REGISTRY2 studies |
| *adhoc* | visit | RET | Ad Hoc data including: Variable, Motor, Function, TFC, MMSE, Cognitive assessment data |

For detailed information on each constituent *form*, please refer to the *Annotated eCRFs*.

The number of participants included in each PDS data *file* is illustrated in the figure below.

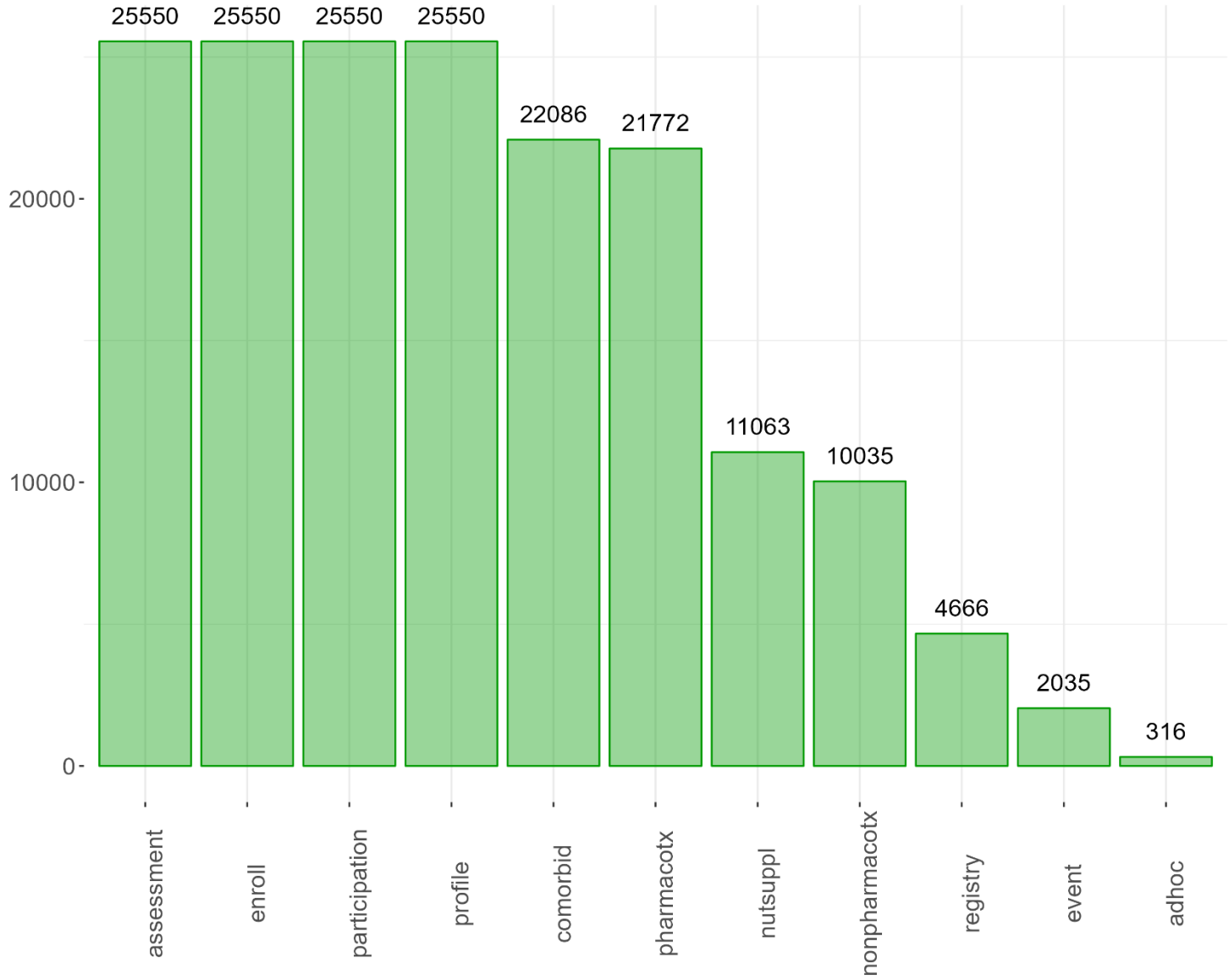


*Figure 1. Number of participants included in each PDS6 data file.*

*Entity relation diagram*

The Enroll-HD periodic dataset data file entity relation diagram is presented in the Figure below. This illustrates the relationship between each of the component data files, along with their key variables (primary keys [PK] and foreign keys [FK])which are required to combine the data files.

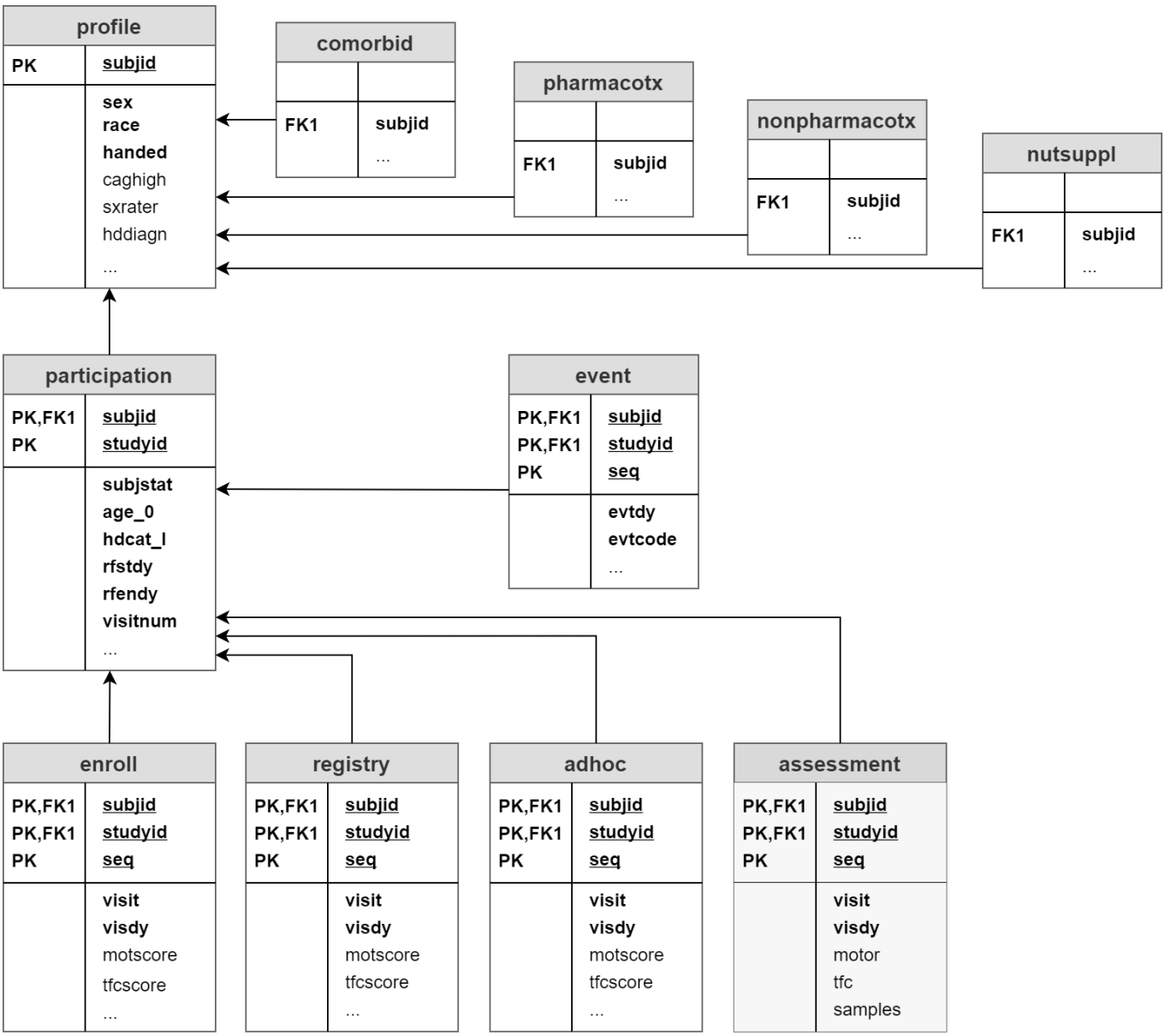


*Figure 2: Entity relationship diagram.*

# Structure of variables

Each Enroll-HD periodic dataset file contains variables. This Data Dictionary lists all variables by form with the appropriate attributes. A list of the attribute types are provided in Table 3.

*Table 3. Data dictionary column fields.*

| Attribute | Description |
|---|---|
| **Label** | Variable label (CDISC SDTM compliant) |
| **Domain** | Assigned CDISC SDTM Domain |
| **Category** | Assigned CDISC SDTM Category (optional) |
| **Form** | Identifies the form where the information is collected during the visit. |
| **Variable** | Internal variable name. Variable is defined in CDISC SDTM compliant naming convention or as close as possible |
| **Data Type** | *Boolean:* Represents the values 1 (yes) and 0 (no).<br><br>*Number:* Represents integer or floating-point data values.<br><br>*Text:* Represents alphanumeric string data values.<br><br>*Date:* The date type is represented as the number of days relative to the date of the participant's Enroll-HD baseline visit date. Note that dates that have been specified in the original data as "incomplete" (e.g., without entry of a day) have been automatically completed by the following rule: use "15" as day if day is missing and use "1" as day and "7" as month if day and month is missing. After the date modification is complete, the number of days relating to the enrollment date is calculated and provided in the dataset. The information about whether a date has been automatically completed is not included in the PDS but can be obtained via SPS request.<br><br>*Single choice:* Variable with assigned list of options where one item can be selected. The value provided in the dataset is taken from the available options list. The list of options is defined as parameters in the data dictionary tables. |
| **Parameter** | Parameter value of coded variables (optional) |
| **Coding** | Internal parameter value of coded variables (optional) |
| **Unit** | Unit of input field (optional) |
| **Transformation** | One important objective of the periodic dataset is to de-identify the Enroll-HD data in order to minimize the possibility to identify a participant. Therefore, many variables are transformed, recoded or outliers removed/cut. These transformations are described on a variable-by-variable basis. |
| **Availability** | All variables in the Enroll-HD dataset are listed in the Data Dictionary. This availability column allows the researcher to identify which variables are available in the PDS ("PDS"), which are available via special request ("available upon SRC approval"), and which are restricted ("not available"). |

# Ordering of visit data

The Enroll-HD PDS data files '*enroll*', '*registry*' and '*adhoc*' contain data for all baseline and follow-up visits, for each participant, for each study. There is not a separate file for each follow-up visit.

The variable *visdy* indicates the timing of each visit in days relative to the Enroll-HD baseline visit date (please refer to the *Date Values* section for further information).

In addition, the files '*enroll*', '*registry*' and '*adhoc*' also include a variable called *seq*. This variable refers to the sequence of the visits and will enable the data analyst to order visits temporally. The *seq* value is in accordance with number of days after the baseline visit (*visdy*), where *seq*=1 refers to the baseline visit, *seq* =2 refers to the 1st follow-up visit, *seq*=3 to the 2nd follow-up visit, and so on, including unscheduled and phone contact visits.

*Table 4. Visit sequencing and visit day example.*

| *subjid* | *studyid* | *visit* | *seq* | *visdy* |
|---|---|---|---|---|
| R000000001 | ENR | Baseline | 1 | 0 |
| R000000001 | R3 | Baseline | 1 | -728 |
| R000000001 | R3 | Follow up | 2 | -363 |

Phone contact visits only occur in the '*enroll*' file. These visits contain missed visit information, reason for missed follow-up visit and participant's availability to continue the study. If these data are not required for your analyses, these visits can be filtered out.

Unscheduled visits occur in the '*enroll*' and '*registry*' files. These visits contain all the same information as a follow-up visit. These visits occur outside the visit window defined. If these data are not required for your analyses, these visits can be filtered out.

# Merging and aligning files

Enroll-HD PDS releases contains one key variable, *subjid*, and it is included in every data file. This allows the user to merge two ormore data files, linking information for each participant across data files.

The key variable *subjid*, labeled as HDID (recoded), is obtained by recoding the HDID. Note that the HDID is a unique participant identifier used across multiple HD studies. HDIDs are not included in any PDS release.

To merge longitudinal data available in visit-based data files, it is important to take into consideration the variable *seq*, as this variable provides information on the visit sequence. Visit day (*visdy*) is also available for sequencing visit data temporally.

**WARNING:** Merging data files in Excel can cause misalignment. Before analyzing the data, check that the resulting merged data file correctly lines up across appropriate fields. To avoid issues with merging data files, it is highly recommended that you use a reputable statistical software package.

Below we provide guidance on selecting entries/lines using Excel or R, respectively. The example described below comprises merging age of HD diagnosis (*hddiagn*) from the *profile* file to *age* at the last visit of each participant in the *enroll* file.

## EXCEL

1. Sort your *enroll* database on the first level by *subjid*, then add a level by *seq* (Smallest to Largest);
2. Create a new column with the name "*select*";
3. On the first row of this column type the formula "`=IF(A2=A3,"","1")`", where `A` corresponds to the column of *subjid* and `A2` to the first row/value of *subjid*. Then press Enter key and drag Auto fill to copy the formula to the range you need. This will create a column with the value "*1*" on the row with maximum *seq* for each participant;
4. Filter for the variable *select* with the value "*1*";
5. Create a new column with the name of the variable you want to merge in '*enroll*' file (in this case *hddiagn*);
6. In that new column, use a VLOOKUP() function to merge hddiagn from profile file, using the variable subjid as a linker. Then press Enter key and drag Auto fill to copy the formula to the range you need.

R

1. Select the rows with highest value in the *seq* variable for each participant

```
install.packages('dplyr')
library(dplyr)
new_database<- enroll %>% group_by(subjid) %>% slice(which.max(seq))
```

2. Use the `merge()` function to merge the variable *hddiagn* from *profile* file. Example:

```
final_database <- merge(new_database, profile[, c("subjid", "hddiagn")],
by="subjid")
```

Note: Each user may use different codes to reach the same results, this is just an exemplar way to perform the task proposed. We recommend the user to read and follow the guidelines available in:

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL https://www.R-project.org/.

# Identifying PDS participants with data from Enroll-HD only

As described above, all individuals in the PDS are Enroll-HD participants. However, a subset of Enroll-HD participants also took part in the REGISTRY study. Should you wish to distinguish between Enroll-HD participants who migrated from the REGISTRY study from those who did not, there are several methods to do so. One simple solution is as follows:

EXCEL:

7. Create a new column with the name "*match_registry*"
8. On the first row of the new variable enter the code:

   `=IF(ISERROR(VLOOKUP(A2,registry.csv!$A:$A,1,FALSE)),"No Match", "Match")`

   where `A2` is the first row of the variable *subjid*, "`registry.csv!$A:$A`" is the column of *subjid* in the *registry* file and `"1"` is the index of the column of *subjid* in the *registry* file. Then press `Enter` key and drag Auto fill to copy the formula to the range you need.
9. The new column *match_registry* will have value *Match* if the participant has migrated from Registry and value *No Match* otherwise. If you are interested only in migrated participants, just filter the variable for the value *Match*.

R:

1. Apply a filter to the *enroll* database

```
library(dplyr)
final_database<- enroll %>%  filter(subjid %in% registry$subjid)
```

   This piece of code will return only the rows of the *enroll* database that belong to the participants migrated from Registry.